A Structured Vector Space Model for Hidden Attribute Meaning in Adjective-Noun Phrases

Matthias Hartung **Anette Frank**
Computational Linguistics Department
Heidelberg University

COLING 2010
Beijing, August 24

Background: Learning Concept Descriptions

- ontology learning: describe and distinguish concepts by properties and relations
  - **motorcycle**: ride, rider, sidecar, park, road, helmet, collision, vehicle, car, moped, ...

    Baroni et al. (2010)
  - **car**: acceleration, performance, front, engine, backseat, chassis, speed, weight, color, condition, driver, buyer, ...

    Poesio & Almuhareb (2005)
- common denominator: learn "prototypical", "static" knowledge about concepts from text corpora

Focus of this Talk

Concept Modification in Linguistic Contexts

- What are the **attributes** of a concept that are highlighted in an adjective-noun phrase ?
- well-known problem in formal semantics: *selective binding*
  - *fast car* ⇔ SPEED(car)=fast
  - *red balloon* ⇔ COLOR(balloon)=red
  - *oval table* ⇔ SHAPE(table)=oval

(cf. Pustejovsky 1995)

- attribute selection as a **compositional process**

Previous Work: Attribute Learning from Adjectives

1. Cimiano (2006):
   - goal: learn binary **noun-attribute relations**
   - detour via adjectives modifying the noun
   - for each adjective: look up attributes from WordNet
2. Almuhareb (2006):
   - goal: learn binary **adjective-attribute relations**
   - pattern-based approach:

```
the ATTR of the * is|was ADJ
```

Problem: The **ternary attribute relation**

$$\textsc{attribute}(\text{noun}) = \text{adjective}$$

is missed by both approaches; e.g.: *hot summer* vs. *hot soup*

Learning Ternary Attribute Relations

"Naive" Solution: Pattern-based Approach

- `the ATTR of the N is|was ADJ`
- challenge: overcome **sparsity issues**

A Structured VSM for Ternary Attribute Relations

- represent adjective and noun meanings independently in a **structured vector space model**
- semantic vectors capture binary relations $r' = \langle noun, attr \rangle$ and $r'' = \langle adj, attr \rangle$
- use **vector composition** to approximate the ternary attribute relation r from r' and r'':

$$v(r) \approx v(r') \otimes v(r'')$$

ex.: $v(\langle speed, car, fast \rangle) \approx v(\langle car, speed \rangle) \otimes v(\langle fast, speed \rangle)$

Outline

Building Vector Representations for Adjectives

	COLOR	DIRECT.	DURAT.	SHAPE	SIZE	SMELL	SPEED	TASTE	TEMP.	WEIGHT
enormous	1	1	0	1	45	0	4	0	0	21

Building Vector Representations for Adjectives

	COLOR	DIRECT.	DURAT.	SHAPE	SIZE	SMELL	SPEED	TASTE	TEMP.	WEIGHT
enormous	1	1	0	1	45	0	4	0	0	21

- 10 manually selected attributes: *color*, *direction*, *duration*, *shape*, *size*, *smell*, *speed*, *taste*, *temperature*, *weight*

  Almuhareb (2006)

- vector component values: raw corpus frequencies obtained from lexico-syntactic patterns

Building Vector Representations for Adjectives

	COLOR	DIRECT.	DURAT.	SHAPE	SIZE	SMELL	SPEED	TASTE	TEMP.	WEIGHT
enormous	1	1	0	1	45	0	4	0	0	21

- 10 manually selected attributes: *color*, *direction*, *duration*, *shape*, *size*, *smell*, *speed*, *taste*, *temperature*, *weight*

  Almuhareb (2006)

- vector component values: raw corpus frequencies obtained from lexico-syntactic patterns

  (A1) `ATTR of DT? NN is|was JJ`
  (A2) `DT? RB? JJ ATTR`
  (A3) `DT? JJ or JJ ATTR`
  (A4) `DT? NN's ATTR is|was JJ`
  (A5) `is|was|are|were JJ in|of ATTR`

Building Vector Representations for Nouns

	COLOR	DIRECT.	DURAT.	SHAPE	SIZE	SMELL	SPEED	TASTE	TEMP.	WEIGHT
enormous	1	1	0	1	45	0	4	0	0	21
ball	14	38	2	20	26	0	45	0	0	20

- 10 manually selected attribute nouns: *color*, *direction*, *duration*, *shape*, *size*, *smell*, *speed*, *taste*, *temperature*, *weight*
- vector component values: raw corpus frequencies obtained from lexico-syntactic patterns

(N1) `NN with|without DT? RB? JJ? ATTR`
(N2) `DT ATTR of DT? RB? JJ? NN`
(N3) `DT NN's RB? JJ? ATTR`
(N4) `NN has|had a|an RB? JJ? ATTR`

Vector Composition

- component-wise multiplication $\odot$
- vector addition $\oplus$

Mitchell & Lapata (2008)

Vector Composition

- component-wise multiplication $\odot$
- vector addition $\oplus$

Mitchell & Lapata (2008)

	COLOR	DIRECT.	DURAT.	SHAPE	SIZE	SMELL	SPEED	TASTE	TEMP.	WEIGHT
enormous	1	1	0	1	45	0	4	0	0	21
ball	14	38	2	20	26	0	45	0	0	20

Vector Composition

- component-wise multiplication $\odot$
- vector addition $\oplus$

Mitchell & Lapata (2008)

	COLOR	DIRECT.	DURAT.	SHAPE	SIZE	SMELL	SPEED	TASTE	TEMP.	WEIGHT
enormous	1	1	0	1	45	0	4	0	0	21
ball	14	38	2	20	26	0	45	0	0	20
enormous $\odot$ *ball*	14	38	0	20	**1170**	0	180	0	0	420
enormous $\oplus$ *ball*	15	39	2	21	**71**	0	49	0	0	41

Vector Composition

- component-wise multiplication $\odot$
- vector addition $\oplus$

Mitchell & Lapata (2008)

	COLOR	DIRECT.	DURAT.	SHAPE	SIZE	SMELL	SPEED	TASTE	TEMP.	WEIGHT
enormous	1	1	0	1	45	0	4	0	0	21
ball	14	38	2	20	26	0	45	0	0	20
enormous $\odot$ *ball*	14	38	0	20	**1170**	0	180	0	0	420
enormous $\oplus$ *ball*	15	39	2	21	**71**	0	49	0	0	41

- **expectation:** vector multiplication comes closest to the linguistic function of intersective adjectives !

Attribute Selection

- goal: make attributes explicit that are most salient in the compositional semantics of adjective-noun phrases
- achieved so far: ranking of attributes according to their prominence in the composed vector representation
- **attribute selection:** distinguish meaningful from noisy components in vector representations
  - MPC Selection
  - Threshold Selection
  - Entropy Selection
  - Median Selection

MPC Selection

Functionality:

- selects the **m**ost **p**rominent **c**omponent from each vector (in terms of absolute frequencies)

	COLOR	DIRECT.	DURAT.	SHAPE	SIZE	SMELL	SPEED	TASTE	TEMP.	WEIGHT
enormous	1	1	0	1	45	0	4	0	0	21

Drawback:

- inappropriate for vectors with more than one meaningful dimension

Threshold Selection

Functionality:

- selects all components exceeding a frequency threshold θ (here: $\theta \geq 10$)

	COLOR	DIRECT.	DURAT.	SHAPE	SIZE	SMELL	SPEED	TASTE	TEMP.	WEIGHT
ball	14	38	2	20	26	0	45	0	0	20

Drawbacks:

- introduces an additional parameter to be optimized
- difficult to apply to composed vectors
- unclear whether method scales to vectors of higher dimensionality

Entropy Selection

Functionality:

- select all **informative components**
- information theory: gain in entropy $\equiv$ loss of information
- retain all (combinations of) components that lead to a gain in entropy when taken out

	COLOR	DIRECT.	DURAT.	SHAPE	SIZE	SMELL	SPEED	TASTE	TEMP.	WEIGHT
enormous	1	1	0	1	45	0	4	0	0	21
ball	14	38	2	20	26	0	45	0	0	20

Drawback:

- yields no attribute for vectors with broad and flat distributions (noun vectors, in particular)

Median Selection

Functionality:

- tailored to noun vectors, in particular
- select all components with values **above the median**

	COLOR	DIRECT.	DURAT.	SHAPE	SIZE	SMELL	SPEED	TASTE	TEMP.	WEIGHT
ball	14	38	2	20	26	0	45	0	0	20

Drawback:

- depends on the number of dimensions

Taking Stock...

Experimental Setup

Experiments:

1. attribute selection from adjective vectors
2. attribute selection from noun vectors
3. attribute selection from composed adjective-noun vectors

Methodology:

- vector acquisition from ukWaC corpus (Baroni et al. 2009)
- gold standards for comparison:
  - Experiment 1: compiled from WordNet
  - Experiments 2/3: manually established by human annotators
- evaluation metrics: precision, recall, f_1-score

Experiment 1: Attribute Selection from Adjective Vectors

Data Set

- all adjectives extracted by patterns (A1)-(A5) occurring at least 5 times in ukWaC (3505 types in total)

Gold Standard

- 1063 adjectives that are linked to at least one of the ten attributes we consider in WordNet 3.0

Baseline: Re-Implementation of Almuhareb (2006)

- patterns (A1)-(A3) only
- manually optimized thresholds for attribute selection
- frequency scores acquired from the web

Experiment 1: Results

	Almuhareb (reconstr.)				VSM (TSel + Target Filter)					VSM (ESel + Target Filter)			
	P	R	F	Thr	P	R	F	Patt	Thr	P	R	F	Patt
A1	0.183	0.005	0.009	5	0.300	0.004	0.007	A3	5	**0.519**	0.035	0.065	A3
A2	0.207	0.039	0.067	50	0.300	0.033	0.059	A1	50	0.240	0.049	0.081	A3
A3	0.382	0.020	0.039	5	0.403	0.014	0.028	A1	5	0.375	0.027	0.050	A1
A4					0.301	0.020	0.036	A3	10	0.272	0.020	0.038	A1
A5					0.295	0.008	0.016	A3	24	0.315	0.024	0.045	A3
all					0.420	0.024	0.046	A1	183	0.225	0.054	0.087	A3

Table: Attribute Selection from Adjective Vectors

- re-implementation yields performance comparable to Almuhareb's original system
- performance increase of 13 points in precision over Almuhareb; recall is still poor
- best parameter settings:
  - entropy selection method
  - target filtering (intersect extractions of two patterns in order to remove noisy or unreliable vectors)

Experiment 2: Attribute Selection from Noun Vectors

Creation of an Annotated Data Set

- random sample from the balanced set of 402 (216) nouns compiled by Almuhareb (2006)
- three human annotators
- task: remove all attributes that are not appropriate for any sense of a given noun
- adjudication of disagreements by majority voting

Resulting Gold Standard

- 100 nouns with 4.24 attributes on average
- inter-annotator agreement: $\kappa = 0.69$

Experiment 2: Results

	MPC			ESel			MSel		
	P	R	F	P	R	F	P	R	F
N1	0.22	0.06	0.10	0.29	0.04	0.07	0.22	0.09	0.13
N2	0.29	0.18	0.23	0.20	0.06	0.09	0.28	0.39	0.33
N3	**0.34**	0.05	0.09	0.20	0.02	0.04	0.25	0.08	0.12
N4	0.25	0.02	0.04	0.29	0.02	0.03	0.26	0.02	0.05
all	0.29	0.18	0.22	0.20	0.06	0.09	0.28	**0.43**	**0.34**

Table: Attribute Selection from Noun Vectors

- MPC: relatively precise, poor in terms of recall
- ESel: counterintuitively fails to increase recall
- MSel: best recall, most suitable for this task

Problems:

- vectors with broad, flat distributions
- binary attribute-noun relation often not overtly realized

Experiment 3: Attribute Selection from Composed Adjective-Noun Vectors

Creation of an Annotated Data Set

- partially random sample from 386 property-denoting adjectives $\times$ 216 nouns
- three human annotators (same as in Experiment 2)
- task: remove all attributes not appropriate for a given pair (not provided by the noun or not selected by the adjective)
- adjudication of disagreements by majority voting

Resulting Gold Standard

- 76 pairs with 1.13 attributes on average, 24 "empty" pairs
- inter-annotator agreement: $\kappa = 0.67$

Experiment 3: Baselines

- **BL-P:** purely **pattern**-based method searching for patterns that make ternary attribute relations explicit

  `the ATTR of the N is|was ADJ`

- **BL-A:** take individual **adjective vector** as surrogate for composition
- **BL-N:** take individual **noun vector** as surrogate for composition

Experiment 3: Results

	MPC			ESel			MSel		
	P	R	F	P	R	F	P	R	F
Adj $\odot$ *N*	0.60	0.58	**0.59**	**0.63**	0.46	0.54	0.27	0.72	0.39
Adj $\oplus$ *N*	0.43	0.55	0.48	0.42	0.51	0.46	0.18	**0.91**	0.30
BL-Adj	0.44	0.60	0.50	0.51	0.63	0.57	0.23	0.83	0.36
BL-N	0.27	0.35	0.31	0.37	0.29	0.32	0.17	0.73	0.27
BL-P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table: Attribute Selection from Composed Adjective-Noun Vectors

- complete failure of BL-P
- modelling ternary relations by composing vector representations of reduced complexity is feasible, but: choice of composition method matters
- ESel most suitable wrt. precision (partly due to its ability to return "empty" selections)
- robustness of MPC mainly due to the large proportion of pairs in the test set that elicit one attribute only

Conclusions and Outlook

- **structured VSM** as a framework for inferring hidden attributes in the compositional semantics of adjective-noun phrases
- **vector composition** as a hinge to model ternary attribute relations from individual vectors capturing adjective and noun meanings, thus **avoiding sparsity issues**
- attribute selection from adjectives: **increase of 13 points** in precision above pattern-based approach of Almuhareb (2006)
- future work:
  - scale approach to higher dimensionality
  - address problems with infrequent and unreliable vectors (particularly nouns)

References

- Almuhareb, Abdulrahman (2006): *Attributes in Lexical Axcquisition*. Ph.D. Thesis, University of Essex.
- Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta (2009): The Wacky Wide Web. A Collection of Very Large Linguistically Processed Web-Crawled Corpora, in: *Journal of Language Resources and Evaluation 43 (3)*: 209-226.
- Baroni, Marco, Brian Murphy, Eduard Barbu & Massimo Poesio (2010): Strudel. A Corpus-based Semantic Model Based On Properties and Types, in: *Cognitive Science 34 (2)*: 222-254.
- Cimiano, Philipp (2006): *Ontology Learning and Population. Algorithms, Evaluation and Applications*. Springer.
- Mitchell, Jeff & Mirella Lapata (2008): *Vector-based Models of Semantic Composition*, in: Proc. of ACL-08/HLT. Columbus, Ohio: 236-244.
- Poesio, Massimo & Abdulrahman Almuhareb (2005): *Identifying Concept Attributes Using a Classifier*, in: Proc. of the ACL Workshop on Frontiers in Corpus Annotation. Ann Arbor, Michigan: 76-83.
- Pustejovsky, James (1995): *The Generative Lexicon*. MIT Press.

Thanks...

...for your attention.
Questions ?